multi-head Attention in Transformers

what is the drawback with the self-attention ?.

there is ambuity in the sentence.

The man saw the astronomer with a telescope

	The	man	saw	the	astronomer	with	a	telescope
The								
man								
saw								
the								
astronomer								
with								
a								
telescope								

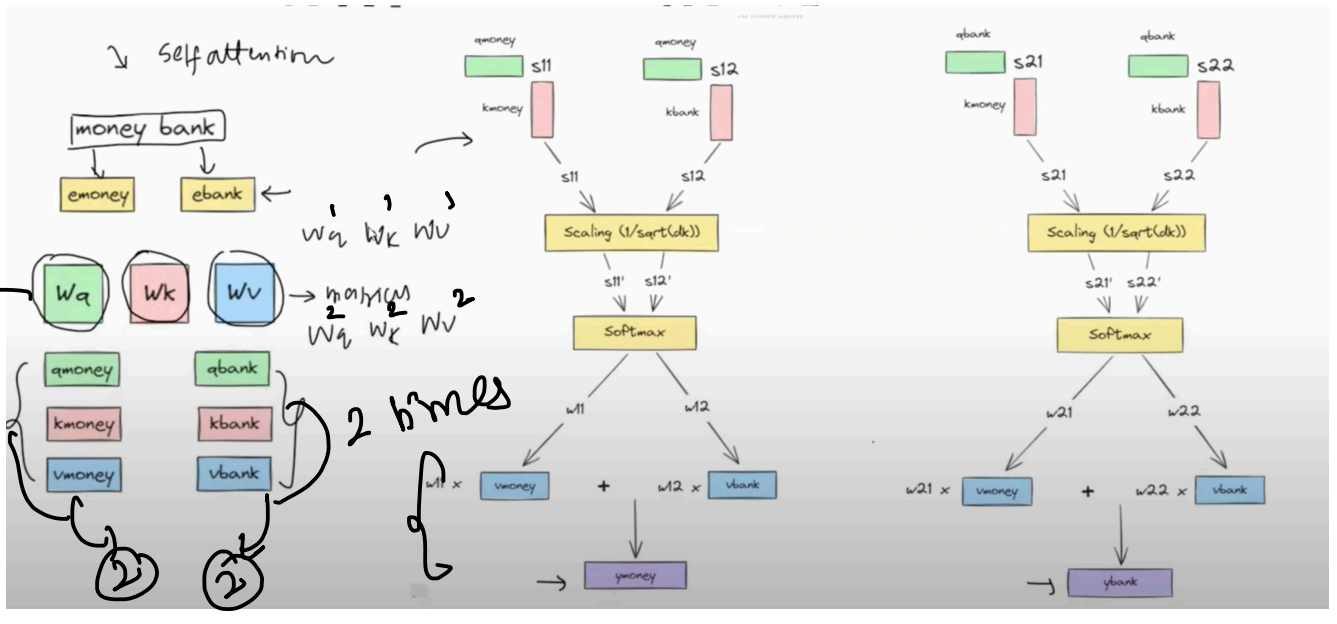
There are 2 meaning to this

telescope

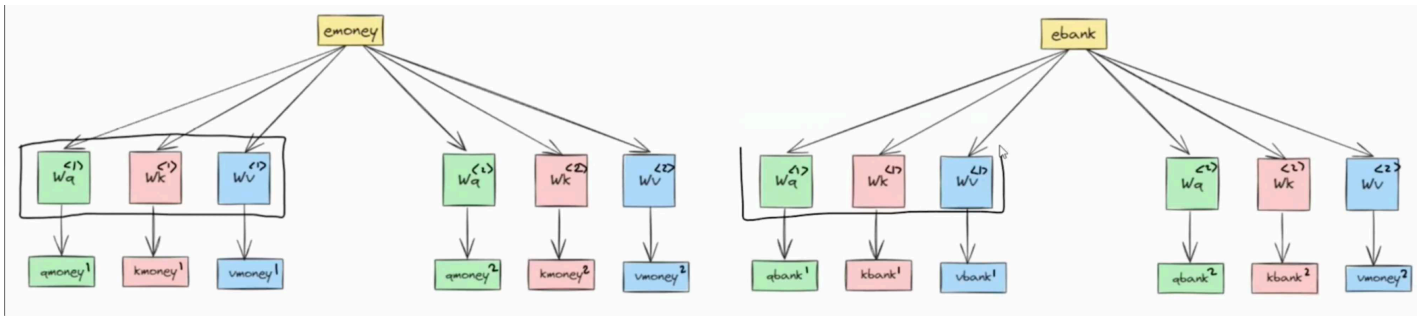
∴ But self-attention can capture only one meaning.

To solve this issue, multi-head attention was introduced.

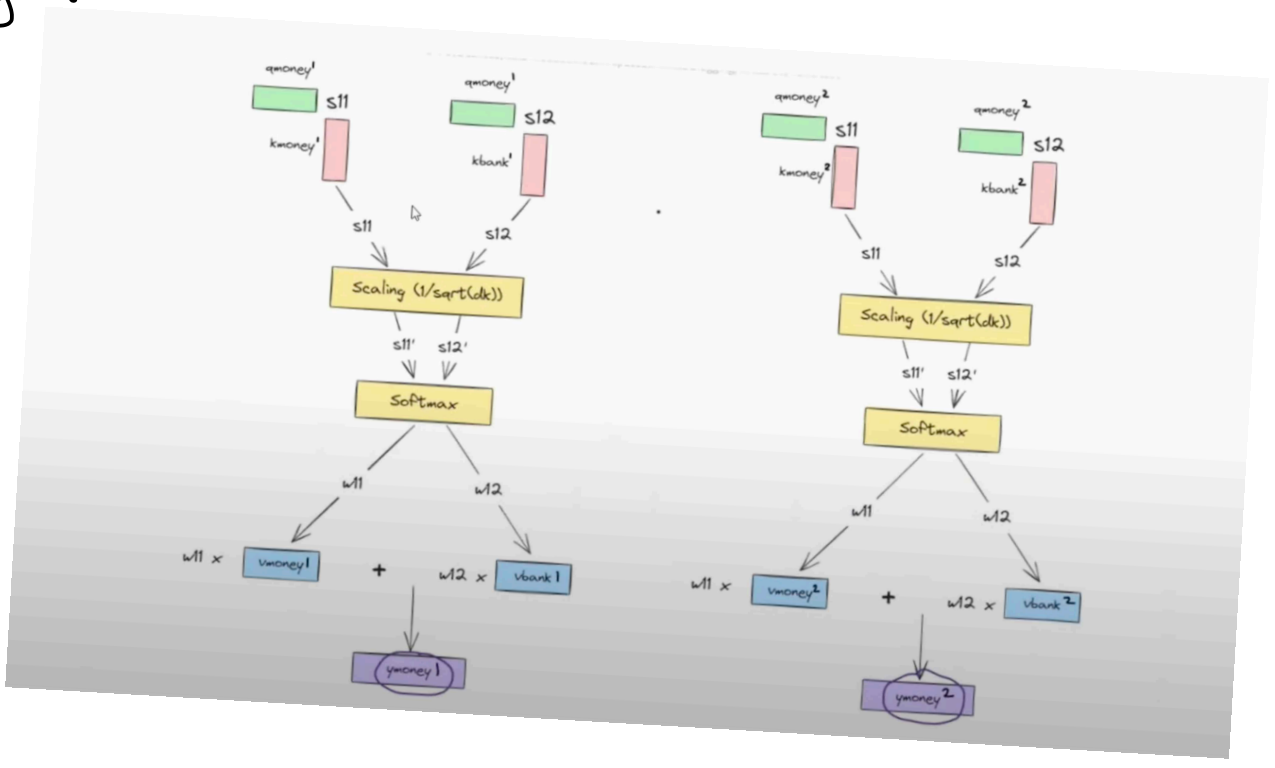
∴ Basically multi-head attention is more than one self-attention. One head is one self-attention.

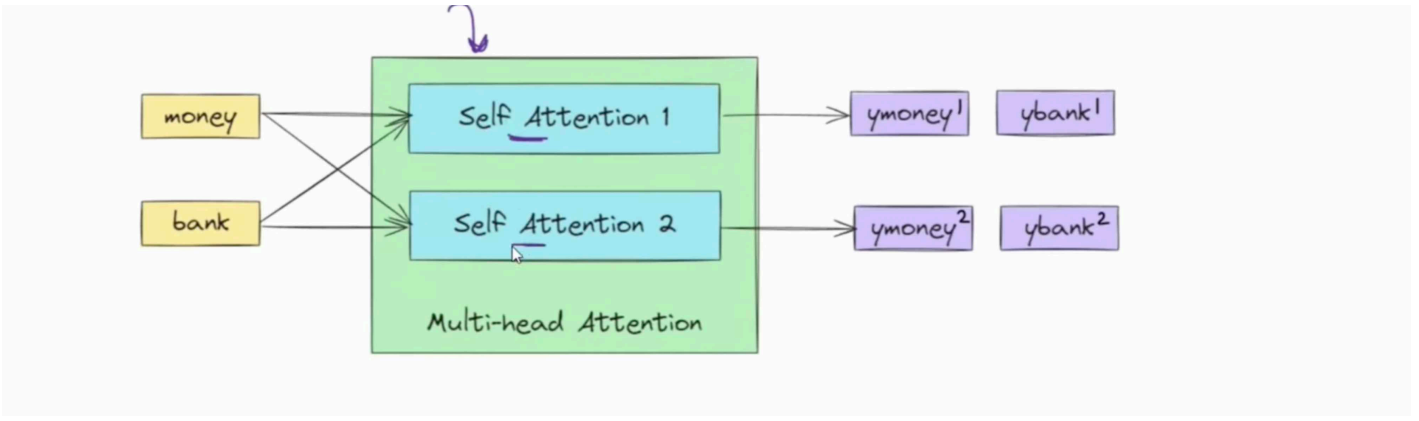


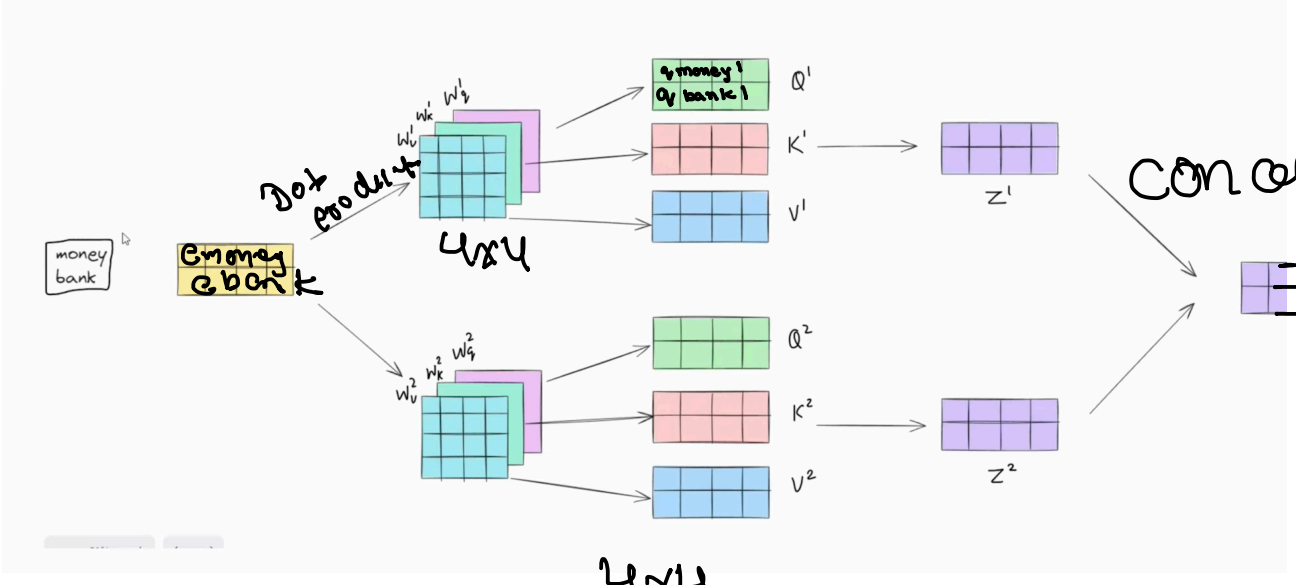
→ When working on self-attention, we have only one set of W_Q, W_K, W_V. But in case of multi-head attention, we have this set more than once.

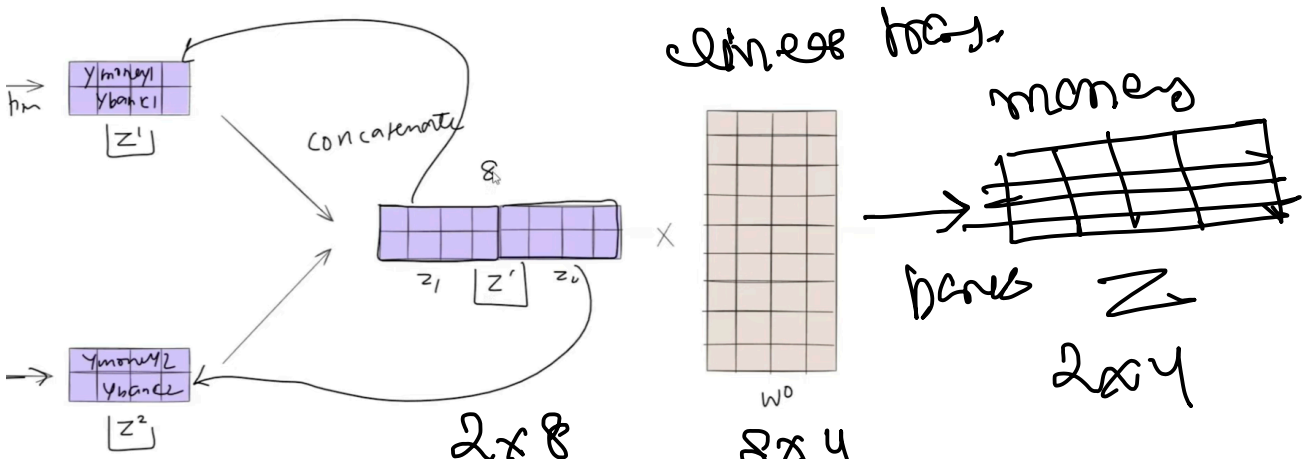


As we see below, 2 ymoney are generated.



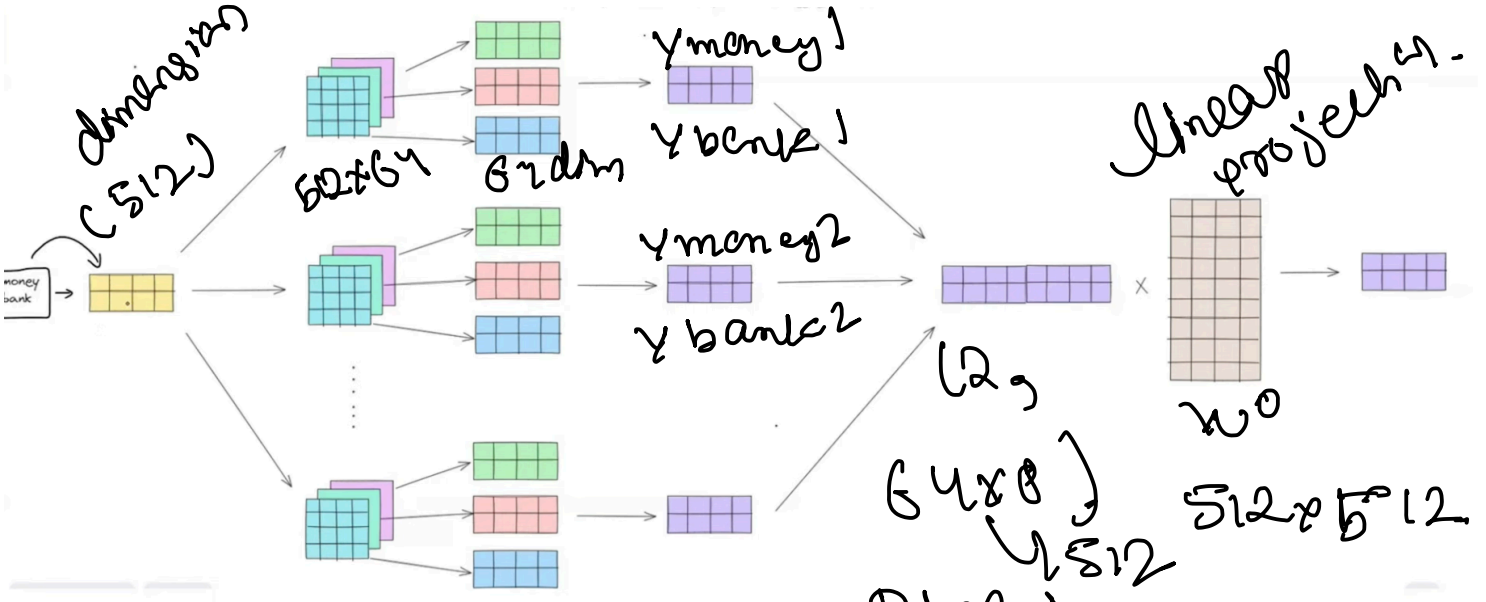

money
bank
Self Attention 1
Self Attention 2
Multi-head Attention
ymoney¹
ybank¹
ymoney²
ybank²


money
bank
Cmoney
Cbank
Dot product
4x4
Q¹
K¹
V¹
Z¹
Q²
K²
V²
Z²
Concatenate
4x4


Concatenate
8
Z1
Z'
Z2
W0
2x8
8x4
2x4
money
bank
Z
2x4

Attention is all you need

Attention



we reduce
the Dimension
Because we wanted to
reduce the computation.